

FORMALE SYSTEME

11. Vorlesung: Von regulären zu kontextfreien Sprachen

Markus Krötzsch
 Lehrstuhl Wissensbasierte Systeme

TU Dresden, 16. November 2017

Die Chomsky-Hierarchie

Die **Chomsky-Hierarchie** unterteilt Grammatiken in vier Stufen:

- **Typ 0:** beliebige Grammatiken
- **Typ 1: kontextsensitive Grammatiken:**
 - (a) Alle Regeln $w \rightarrow v$ erfüllen die Bedingung $|w| \leq |v|$.
 - (b) Es gibt eine Regel $S \rightarrow \epsilon$ und alle anderen Regeln $w \rightarrow v$ enthalten kein S in v und erfüllen $|w| \leq |v|$.
- **Typ 2: kontextfreie Grammatiken:**
Alle Regeln haben die Form $A \rightarrow v$ für eine Variable A .
- **Typ 3: reguläre Grammatiken:**
Alle Regeln haben eine der Formen

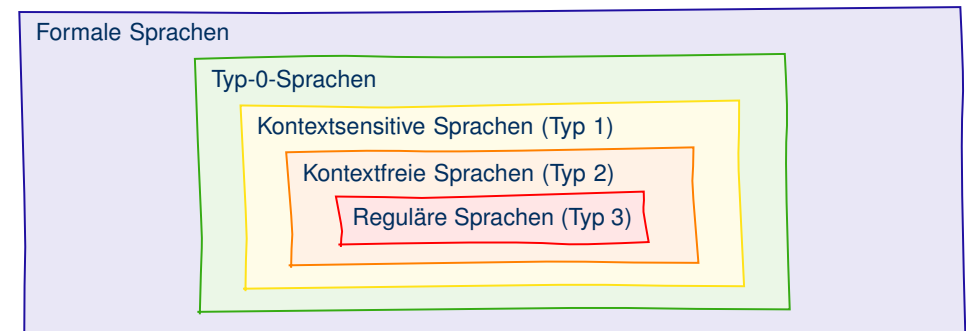
$$A \rightarrow cB \quad A \rightarrow c \quad A \rightarrow \epsilon$$

wobei A und B Variablen sind und c ein Terminalsymbol ist.

Viereinhalb Wochen reguläre Sprachen

auf sieben Folien

Chomsky's Hierarchie ist eine Hierarchie



(Dafür mussten wir Typ-1 erweitern und ϵ -Regeln bei Typ-2 eliminieren.)

Automaten

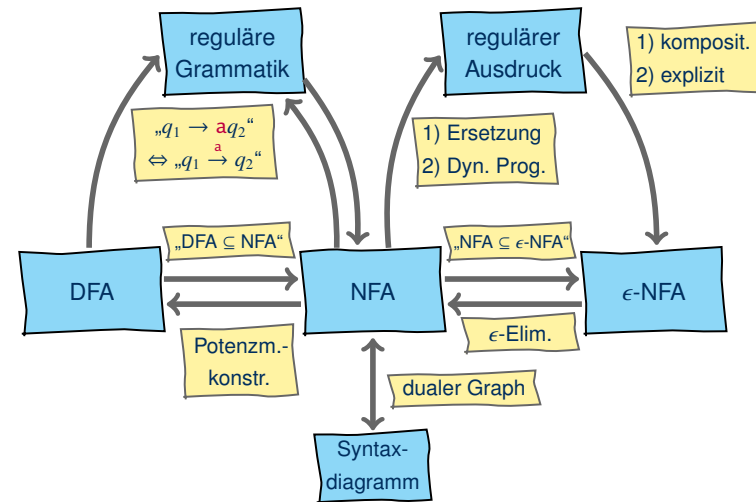
Wir kennen mehrere Varianten endlicher Automaten:

- **Deterministischer endlicher Automat (DFA)**
 - mit totaler Übergangsfunktion
- **Nichtdeterministischer endlicher Automat (NFA)**
 - mit ϵ -Übergängen
 - mit Wortübergängen
 - mit Übergängen für reguläre Ausdrücke (nur für Umwandlung reg. Ausdruck $\rightarrow \epsilon$ -NFA)

Die Sprache eines Automaten haben wir auf zwei Arten definiert

- Mithilfe einer verallgemeinerten Übergangsfunktion, die ganze Wörter einliest
- Durch akzeptierende Läufe, die einem Wort zugeordnet werden können

Darstellungen von Typ-3-Sprachen



Umformungsalgorithmen (1)

Eingabe	Ausgabe	Vorlesung
kontextfreie Grammatik	ϵ -freie kontextfreie Grammatik	2
DFA \mathcal{M}	totaler DFA $\mathcal{M}_{\text{total}}$	3
DFA \mathcal{M}	reguläre Grammatik $G_{\mathcal{M}}$	3
Syntaxdiagramm	NFA	4
NFA \mathcal{M}	Potenzmengen-DFA \mathcal{M}_{DFA}	4
reguläre Grammatik G	NFA \mathcal{M}_G	5
NFA mit Wortübergängen	ϵ -NFA	5
ϵ -NFA \mathcal{M}	NFA $\text{elim}_{\epsilon}(\mathcal{M})$	5

Umformungsalgorithmen (2)

Eingabe	Ausgabe	Vorlesung
NFAs $\mathcal{M}_1, \mathcal{M}_2$	Vereinigungs-NFA $\mathcal{M}_1 \oplus \mathcal{M}_2$	5
NFAs/DFAs $\mathcal{M}_1, \mathcal{M}_2$	Produkt-NFA/DFA $\mathcal{M}_1 \otimes \mathcal{M}_2$	5
totaler DFA \mathcal{M}	Komplement-DFA $\overline{\mathcal{M}}$	5
NFAs $\mathcal{M}_1, \mathcal{M}_2$	ϵ -NFA $\mathcal{M}_1 \odot \mathcal{M}_2$ für Konkatenation	5
NFA \mathcal{M}	ϵ -NFA \mathcal{M}^* für Kleene-Abschluss	5
regulärer Ausdruck	ϵ -NFA (Komposition)	6
regulärer Ausdruck	ϵ -NFA (explizit)	6
NFA	regulärer Ausdruck (Gleichungssystem)	7
NFA	regulärer Ausdruck (dyn. Programmierung)	7
totaler DFA \mathcal{M}	Quotienten-DFA \mathcal{M}/\sim	8
totaler DFA \mathcal{M}	reduzierter DFA \mathcal{M}_r	8

Reguläre Sprachen

Die Menge der regulären Sprachen ist ...

- die Menge genau all jener Sprachen ...
 - die von einer Typ-3-Grammatik beschrieben werden
 - die von einem DFA erkannt werden
 - die von einem NFA erkannt werden
 - die durch einen regulären Ausdruck beschrieben werden
 - die endlich viele Myhill-Nerode-Kongruenzklassen haben
- die kleinste Menge von Sprachen ...
 - welche alle endlichen Sprachen enthält und unter \cap , \cup , $\bar{}$, \circ und $*$ abgeschlossen ist
 - welche die Sprachen \emptyset , $\{\epsilon\}$ und $\{a\}$ ($a \in \Sigma$) enthält und unter \cup , \circ und $*$ abgeschlossen ist

Alle endlichen Sprachen sind regulär (aber nicht umgekehrt)

Alle regulären Sprachen erlauben Pumping (aber nicht umgekehrt)

Probleme für endliche Automaten

Problem	Fragestellung	Komplexität
Leerheit	$L(\mathcal{M}) \stackrel{?}{=} \emptyset$	polynomiell
Inklusion	$L(\mathcal{M}_1) \stackrel{?}{\subseteq} L(\mathcal{M}_2)$	polynomiell falls \mathcal{M}_2 DFA exponentiell falls \mathcal{M}_2 NFA
Äquivalenz	$L(\mathcal{M}_1) \stackrel{?}{=} L(\mathcal{M}_2)$	polynomiell falls \mathcal{M}_1 und \mathcal{M}_2 DFA exponentiell falls \mathcal{M}_1 oder \mathcal{M}_2 NFA
Wortproblem	$w \stackrel{?}{\in} L(\mathcal{M})$	polynomiell
Universalität	$L(\mathcal{M}) \stackrel{?}{=} \Sigma^*$	polynomiell falls \mathcal{M} DFA exponentiell falls \mathcal{M} NFA
Endlichkeit	$L(\mathcal{M})$ endlich?	polynomiell

Kontextfreie Sprachen

Wir hatten kontextfreie Sprachen wie folgt definiert:

Eine **kontextfreie Grammatik** (oder **Typ-2-Grammatik** oder **CFG**) enthält nur Regeln der Form $A \rightarrow v$, wobei A eine Variable ist.

Eine Sprache ist **kontextfrei** (oder **Typ 2**), wenn sie durch eine kontextfreie Grammatik dargestellt werden kann.

Das genügt, um nichtreguläre Sprachen darzustellen:

Beispiel: Die Sprache $\{a^n b^n \mid n \geq 0\}$ ist kontextfrei, da sie durch die folgende CFG dargestellt werden kann:

$$S \rightarrow \epsilon \mid aSb$$

(Übung: Beweise, dass die Grammatik wirklich diese Sprache darstellt.)

Beispiel

Die Grammatik

$$\begin{aligned} S &\rightarrow A \mid M \mid V \\ A &\rightarrow (S+S) \\ M &\rightarrow (S*S) \\ V &\rightarrow x \mid y \mid z \end{aligned}$$

erzeugt zum Beispiel das Wort $(x * (y + z))$ über die Ableitung:

$$\begin{aligned} S &\Rightarrow M \Rightarrow (S*S) \Rightarrow (V*S) \Rightarrow (x*S) \Rightarrow (x*A) \Rightarrow (x*(S+S)) \\ &\Rightarrow (x*(V+S)) \Rightarrow (x*(y+S)) \Rightarrow (x*(y+V)) \Rightarrow (x*(y+z)) \end{aligned}$$

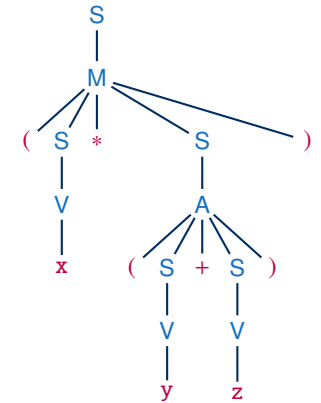
Ableitungen als Bäume

Grammatik:

$$\begin{aligned} S &\rightarrow A \mid M \mid V & A &\rightarrow (S+S) \\ M &\rightarrow (S*S) & V &\rightarrow x \mid y \mid z \end{aligned}$$

Ableitung:

$$\begin{aligned} S &\Rightarrow M \Rightarrow (S*S) \Rightarrow (V*S) \\ &\Rightarrow (x*S) \Rightarrow (x*A) \\ &\Rightarrow (x*(S+S)) \Rightarrow (x*(V+S)) \\ &\Rightarrow (x*(y+S)) \Rightarrow (x*(y+V)) \\ &\Rightarrow (x*(y+z)) \end{aligned}$$



Von Ableitung zu Ableitungsbaum

Sei $G = \langle V, \Sigma, P, S \rangle$ eine Grammatik und sei $S = w_0 \Rightarrow w_1 \Rightarrow \dots \Rightarrow w_n$ eine Ableitung (mit $w_i \in (V \cup \Sigma)^*$ für alle $i \in \{1, \dots, n\}$).

Wir erhalten den entsprechenden **Ableitungsbaum** wie folgt:

- Der Ableitungsbaum wird initialisiert mit einem einzigen Wurzelknoten S
- Der Baum wird schrittweise konstruiert. Nach i Schritten ergeben die Blätter des Baumes – gelesen von links nach rechts – immer genau w_i .
- Wenn in einem Ableitungsschritt $w_i \Rightarrow w_{i+1}$ die Regel $V \rightarrow u$ angewendet wurde, dann erhält der Knoten für V genau $|u|$ Kindknoten, die – von links nach rechts – mit den Symbolen aus u beschriftet werden.

Ableitungsbäume sind auch als **Syntaxbäume** oder **Parsebäume** bekannt

Vom Ableitungsbaum zur Ableitung?

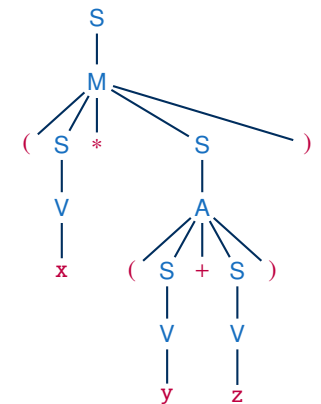
Der selbe Ableitungsbaum wird oft durch viele Ableitungen erzeugt:

Vorige Ableitung:

$$\begin{aligned} S &\Rightarrow M \Rightarrow (S*S) \Rightarrow (V*S) \\ &\Rightarrow (x*S) \Rightarrow (x*A) \\ &\Rightarrow (x*(S+S)) \Rightarrow (x*(V+S)) \\ &\Rightarrow (x*(y+S)) \Rightarrow (x*(y+V)) \\ &\Rightarrow (x*(y+z)) \end{aligned}$$

Alternative Ableitung:

$$\begin{aligned} S &\Rightarrow M \Rightarrow (S*S) \Rightarrow (S*A) \\ &\Rightarrow (S*(S+S)) \Rightarrow (S*(S+V)) \\ &\Rightarrow (S*(V+V)) \Rightarrow (V*(V+V)) \\ &\Rightarrow (V*(V+z)) \Rightarrow (V*(y+z)) \\ &\Rightarrow (x*(y+z)) \end{aligned}$$



Vom Ableitungsbaum zur Ableitung

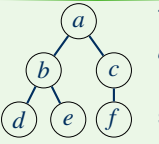
Beobachtung:

- Für jeden inneren Knoten im Ableitungsbaum gibt es genau einen Ableitungsschritt
- Die Reihenfolge der Schritte ist egal, sofern Elternknoten vor ihren Kindern ersetzt werden

Eine totale Ordnung der Knoten eines Baums, bei der Eltern vor ihren Kindern betrachtet werden, heißt **topologische Sortierung**.

~> Jede topologische Sortierung der Knoten eines Ableitungsbaumes führt zu einer erlaubten Ableitung

Beispiel:



Topologische Sortierungen:
abcdef (Breitensuche von links), *acbfd* (Breitensuche von rechts), *abdecf* (Tiefensuche von links), *acfbed* (Tiefensuche von rechts), ...

Rechtsableitungen und Linksableitungen

Man kann diese speziellen Ableitungen auch ohne den Ableitungsbaum direkt erzeugen:

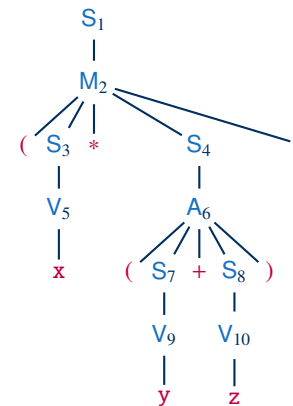
- **Linksableitung:** In jedem Ableitungsschritt wird die am weitesten links stehende Variable ersetzt
- **Rechtsableitung:** In jedem Ableitungsschritt wird die am weitesten rechts stehende Variable ersetzt

Bei CFGs kann jede dieser Strategien jedes erzeugbare Wort generieren

(bei Typ-1-Grammatiken im Allgemeinen nicht – Übung: warum?)

Vom Ableitungsbaum zur Ableitung: Beispiel

Wir markieren die Variablen zur Veranschaulichung mit Indizes:



Sortierung $S_1 M_2 S_3 V_5 S_4 A_6 S_7 V_9 S_8 V_{10}$:

$$\begin{aligned}
 S_1 &\Rightarrow M_2 \Rightarrow (S_3 * S_4) \Rightarrow (V_5 * S_4) \\
 &\Rightarrow (x * S_4) \Rightarrow (x * A_6) \\
 &\Rightarrow (x * (S_7 + S_8)) \Rightarrow (x * (V_9 + S_8)) \\
 &\Rightarrow (x * (y + S_8)) \Rightarrow (x * (y + V_{10})) \\
 &\Rightarrow (x * (y + z))
 \end{aligned}$$

Entspricht Tiefensuche von links

~> **Linksableitung**

Alternative Reihenfolge bei Tiefensuche von rechts:

$S_1 M_2 S_4 A_6 S_8 V_{10} S_7 V_9 S_3 V_5$

~> **Rechtsableitung**

Anwendung Ableitungsbaum

Der Ableitungsbaum ist von großer praktischer Bedeutung, da er die „innere Struktur“ eines Wortes einer kontextfreien Sprache repräsentiert

In der Praxis geht es meist nicht darum, zu prüfen, ob ein Wort in einer Sprache liegt, sondern darum, seine syntaktische Struktur zu ermitteln

Beispiele:

- Parsebäume in der **Verarbeitung natürlicher Sprache** können Aufschluss über die Bedeutung eines Satzes geben
- Syntaxbäume in **Programmiersprachen** sind die Grundlage zur inhaltlichen Interpretation des Codes
- Ableitungsbäume in **Mark-Up-Sprachen** wie HTML oder XML sind entscheidend für die Adressierung von Elementen („DOM-Tree“)

Zusammenfassung und Ausblick

Wir kennen **viele Charakterisierungen für reguläre Sprachen**, die man mit zahlreichen Umformungen in Beziehung setzen kann

Wörter in **kontextfreien Sprachen** haben eine interessante innere Struktur, die wir durch **Ableitungsbäume** darstellen können

Bei Typ-2-Grammatiken repräsentieren Ableitungsbäume mehrere mögliche Ableitungen.

Offene Fragen:

- Wie kann das Wortproblem bei kontextfreien Grammatiken gelöst werden?
- Haben kontextfreie Sprachen ein Berechnungsmodell?
- Wie sehen nicht-kontextfreie Sprachen aus und wie erkennt man sie?